

## PEER REVIEW HISTORY

BMJ Open publishes all reviews undertaken for accepted manuscripts. Reviewers are asked to complete a checklist review form (<http://bmjopen.bmj.com/site/about/resources/checklist.pdf>) and are provided with free text boxes to elaborate on their assessment. These free text comments are reproduced below.

### ARTICLE DETAILS

<b>TITLE (PROVISIONAL)</b>	CRITICAL APPRAISAL AND ISSUES REGARDING GENERALISABILITY OF COMPARATIVE EFFECTIVENESS STUDIES OF NOACS IN ATRIAL FIBRILLATION AND THEIR RELATION TO CLINICAL TRIAL DATA - A SYSTEMATIC REVIEW
<b>AUTHORS</b>	Bunge, Eveline; van Hout, Ben; Haas, Sylvia; Spentzouris, Georgios; Cohen, Alexander

### VERSION 1 – REVIEW

<b>REVIEWER</b>	Ole-Christian Rutherford Oslo University Hospital Ullevål, Oslo, Norway. Institute of Clinical Medicine, University of Oslo, Oslo, Norway.
<b>REVIEW RETURNED</b>	18-Aug-2020

<b>GENERAL COMMENTS</b>	<p>Bunge and colleagues have produced highly relevant research. The manuscript is well written, the language is clear and concise. As a researcher in this field myself I am only too aware of the inherent weaknesses of observational cohort studies. But they are all we have at the moment.</p> <p>A few thoughts or points that I would wish to see more clearly written in the text:</p> <ul style="list-style-type: none"><li>- I do not believe that one single statistical method (multivariate Cox PH Reg, competing risk regression, PSM or IPTW) has been proven superior or more reliable than the others. It could be clearer in the text that there is no "gold standard" here.</li><li>- Observational studies of NOACs are for hypothesis generation only, not suited to make causal inference. Channelling bias is mentioned, but I suggest the authors make it much clearer what lack of randomisation really means for the overall bias of the results. There are very important temporal trends and differences in prescription patterns that greatly influence outcomes, and especially in these first five years of the NOAC era that almost all the studies included in this systematic review are from. That means that channelling bias will shift between different NOACs during the course of the study. It is of importance to know how each study deals with temporal trends.</li></ul> <p>Furthermore, it should be clearer that almost all studies included in this review not only are from similar registries; but also have the same missing values (almost none contain laboratory results such as haemoglobin levels, creatinine, troponins, BNP og liver enzymes; almost none contain information on smoking status or body mass index; of frailty or of lifestyle and diet. The huge amount of missing important variables making appropriate adjustment or matching quite impossible should be stressed.</p>
-------------------------	--

	<p>-More focus should be on the rigidity and quality of sensitivity analyses in each study(as-treated, intention to treat, longer or shorter follow-up time, using different statistical techniques)  - I do not see a discussion of how to properly select variables for which to adjust or match, and whether the studies included have done this. The use of directed acyclic graphs (DAGs) could be mentioned.</p> <p>Over all a well-written highly relevant manuscript.</p>
--	---

<b>REVIEWER</b>	Boyoung Joung Yonsei University
<b>REVIEW RETURNED</b>	25-Aug-2020

<b>GENERAL COMMENTS</b>	<p>This study appraise the published comparative effectiveness studies on non-vitamin K antagonist oral anticoagulants (NOAC) in nonvalvular atrial fibrillation (NVAF). The author concluded that meaningful comparisons between NOACs on the basis of observational data, even after correction for baseline characteristics, may not be reliable due to unmeasured confounders, channelling bias and insufficient sample size.</p> <p>1. "These studies have some important limitations, even though the larger part of the studies were well conducted technically." It would be helpful if this study show the limitation of each study.</p> <p>2. "On the basis of trial results, expected differences are small and a naïve analysis suggests trials with between 7,700 and 59,500 patients are needed to confirm the observed differences in bleedings and between 51,800 and 7,994,300 to confirm differences in efficacy." This result is interesting. It would be better to show the result with Table, and to show how many patients are need to compare each NOAC.</p> <p>3. It would be helpful if this study show the examples of unmeasured cofounder and channeling bias.</p>
-------------------------	--

<b>REVIEWER</b>	Maharaj Singh Marquette University USA
<b>REVIEW RETURNED</b>	19-Nov-2020

<b>GENERAL COMMENTS</b>	<p>The study discussed methodological issues related to Cox Ph and for PS matching used in the sampled studies.  Adding meta-analysis might have added more to result section, which in turn might have further strengthen discussion and conclusion section of the manuscript.  Authors should address the clinical relevance and/or clinical implications from the systematic review and from the analysis of the results.  Furthermore, the authors should provide some suggestions for future research in the areas discussed in this review.</p>
-------------------------	---

## VERSION 1 – AUTHOR RESPONSE

Reviewer #1:

Bunge and colleagues have produced highly relevant research. The manuscript is well written, the language is clear and concise. As a researcher in this field myself I am only too aware of the inherent weaknesses of observational cohort studies. But they are all we have at the moment.

A few thoughts or points that I would wish to see more clearly written in the text:

I do not believe that one single statistical method (multivariate Cox PH Reg, competing risk regression, PSM or IPTW) has been proven superior or more reliable than the others. It could be clearer in the text that there is no "gold standard" here.

We agree with the reviewer that there is no gold standard statistical method. We made this more clear in the text by adding the following sentence in the introduction on page 5, third paragraph, first sentence: , but there is no gold standard and all methods have advantages and disadvantages.

Observational studies of NOACs are for hypothesis generation only, not suited to make causal inference. Channelling bias is mentioned, but I suggest the authors make it much clearer what lack of randomisation really means for the overall bias of the results.

Unfortunately, we cannot predict what the overall bias in results is due to lack of randomization. The magnitude as well as the direction of the effect towards one NOAC or the other remains unknown. We agree that this can be more clearly described, and therefore added the following sentence to the third paragraph of the discussion section: However, it is unknown what the magnitude and direction (i.e. will the differences in effectiveness and safety between NOACs be smaller or larger) of this potential bias due to lack of randomization would be.

There are very important temporal trends and differences in prescription patterns that greatly influence outcomes, and especially in these first five years of the NOAC era that almost all the studies included in this systematic review are from. That means that channeling bias will shift between different NOACs during the course of the study. It is of importance to know how each study deals with temporal trends.

We added the following text to the last sentences of the sixth paragraph of the discussion: , 'Channelling bias therefore likely occurs and might shift between the NOACs. Although in a few studies it was mentioned that selective prescriptions were noticed and that these might have changed over time, none of the included studies dealt with temporal trends in prescription patterns.

Furthermore, it should be clearer that almost all studies included in this review not only are from similar registries; but also have the same missing values (almost none contain laboratory results such as haemoglobin levels, creatinine, troponins, BNP or liver enzymes; almost none contain information on smoking status or body mass index; of frailty or of lifestyle and diet. The huge amount of missing important variables making appropriate adjustment or matching quite impossible should be stressed. We felt that creatine clearance was especially important to mention as a missing value in all studies, as this plays an important role as a potential effect modifier. This was described in the third paragraph of the discussion section. We agree that this is not the only missing value that might be important, therefore we added the following text to that paragraph: Hardly any laboratory results and lifestyle information were included, such as body-mass index, smoking status and alcohol consumption, which are also risk factors for ischaemic stroke and bleeding events respectively..

More focus should be on the rigidity and quality of sensitivity analyses in each study(as-treated, intention to treat, longer or shorter follow-up time, using different statistical techniques)

We agree with the reviewer that these topics are important, but we feel that for the aim of our study, sensitivity analyses to assess residual confounding, which existence was acknowledged in all

included studies, is the most important one to highlight here. We therefore added the following text to the methods section, critical appraisal section: Sensitivity analyses to further explore the magnitude of residual confounding (i.e. case-crossover study designs; clinical details in a subsample; proxy measures; or instrumental variable (IV) techniques)

We added the following text to the results section, before the paragraph on study results:  
Sensitivity analyses

Although in some articles sensitivity analyses were done, none of the included studies further explored the magnitude of residual confounding in their sensitivity analyses using one of the approaches recommended by IPSOR (see methods section).

We added to the discussion section the following text at the end of the third paragraph: The magnitude of residual confounding was not further explored in the sensitivity of the included studies.

I do not see a discussion of how to properly select variables for which to adjust or match, and whether the studies included have done this. The use of directed acyclic graphs (DAGs) could be mentioned. In the results section on page 26, we described if the included studies provided a rationale for the selection of variables, and if so, what they did. To address the point raised by the reviewer, we added the following text to the discussion section, third paragraph, after the comma in the second sentence: in most studies it was not described how the covariates were selected. The ISPOR Good Research Practices for Retrospective Database Analysis recommends to include all factors that are theoretically related to outcome or treatment selection, even if the relation is weak or statistically non-significant. Directed acyclic graphs might be helpful as well.

Over all a well-written highly relevant manuscript.

#### Reviewer #2

This study appraise the published comparative effectiveness studies on non-vitamin K antagonist oral anticoagulants (NOAC) in nonvalvular atrial fibrillation (NVAf). The author concluded that meaningful comparisons between NOACs on the basis of observational data, even after correction for baseline characteristics, may not be reliable due to unmeasured confounders, channelling bias and insufficient sample size.

1. "These studies have some important limitations, even though the larger part of the studies were well conducted technically." It would be helpful if this study show the limitation of each study. With this sentence, we meant to describe that with the data available, the studies were conducted as best as they could have been conducted. But all studies suffer from the same limitation that they cannot adjust for all variables needed. E.g. creatinine clearance seems to be an important effect modifier, but data on creatinine clearance is hardly available. The methods of each study is described in table 1-5, organized per type of method, and the limitations of each method are summarized in the results section. To further clarify, we rewrote the first sentence of the now pre-final paragraph of the discussion into: In conclusion, even though the larger part of these studies are conducted as well as possible considering what data are available, there are some important limitations regarding the generalisability of the study results especially given the relatively high patient number required for a meaningful comparison between NOACs.

2. "On the basis of trial results, expected differences are small and a naïve analysis suggests trials with between 7,700 and 59,500 patients are needed to confirm the observed differences in bleedings and between 51,800 and 7,994,300 to confirm differences in efficacy." This result is interesting. It would be better to show the result with Table, and to show how many patients are need to compare each NOAC.

This naïve analysis was meant to be descriptive and put the sample sizes of the included studies in context with what is actually needed for a study to be sufficiently powered to detect any differences. As this analysis is not meant to be more than a naïve analysis, we decided to only provide the ranges

of the smallest sample size needed between the two NOACs that are expected to have the largest difference in effectiveness and safety, and the largest sample size needed between the two NOACs that are expected to have the smallest difference in effectiveness and safety.

3. It would be helpful if this study show the examples of unmeasured cofounder and channeling bias. We agree with the reviewer and we added text to the discussion, please see our responses to reviewer #1 on these topics.

**Reviewer #3**

The study discussed methodological issues related to Cox Ph and for PS matching used in the sampled studies. Adding meta-analysis might have added more to result section, which in turn might have further strengthen discussion and conclusion section of the manuscript.

We agree and understand that meta-analysis might strengthen discussion and conclusions if a review is focused on having an estimate of the overall results. This was however not the aim of our study. We aimed to have a critical look at the published comparative effectiveness research regarding the methodologies used. Beside this, we will not be able to address the differences between the studies and will not help in having less unmeasured confounding and channeling bias.

Authors should address the clinical relevance and/or clinical implications from the systematic review and from the analysis of the results.

We feel that we describe the implications for clinical practice in the last two sentences of the last paragraph of the manuscript. To better emphasize this, we added before the second final sentence the following wording: In clinical practice....

Furthermore, the authors should provide some suggestions for future research in the areas discussed in this review.

We agree with the reviewer that this would add to the manuscript. Therefore, we included the following sentences to the prefinal paragraph of the discussion section: Ideally, head-to-head trials should be conducted to compare the efficacy/effectiveness and safety of the four NOACs to overcome the methodological issues in the comparative effectiveness studies. To our knowledge, one head to head trial including all four NOACs is currently running. This nationwide cluster randomized cross-over study aims to compare efficacy and safety of the four NOACs and should be ready September 2021 (clinicaltrials.gov; NCT03129490).

**VERSION 2 – REVIEW**

<b>REVIEWER</b>	Boyoung Joung Yonsei University
<b>REVIEW RETURNED</b>	16-Jan-2021
<b>GENERAL COMMENTS</b>	I have no further comment.